Krylov subspace method for evaluating the self-energy matrices in electron transport calculations

Sørensen, Hans Henrik Brandenborg; Hansen, Per Christian; Petersen, D. E.; Skelboe, S.; Stokbro, Kurt

Published in:
Physical Review B Condensed Matter

Link to article, DOI:
10.1103/PhysRevB.77.155301

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Krylov subspace method for evaluating the self-energy matrices in electron transport calculations

Hans Henrik B. Sørensen* and Per Christian Hansen
Department of Informatics and Mathematical Modelling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark

Dan Erik Petersen and Stig Skelboe
Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark

Kurt Stokbro
Nano-Science Center, University of Copenhagen, Universitetsparken 5, Building D, DK-2100 Copenhagen, Denmark
(Received 29 August 2007; revised manuscript received 4 March 2008; published 1 April 2008; corrected 3 April 2008)

We present a Krylov subspace method for evaluating the self-energy matrices used in the Green's function formulation of electron transport in nanoscale devices. A procedure based on the Arnoldi method is employed to obtain solutions of the quadratic eigenvalue problem associated with the infinite layered systems of the electrodes. One complex and two real shift-and-invert transformations are adopted to select interior eigenpairs with complex eigenvalues on or in the vicinity of the unit circle that correspond to the propagating and evanescent modes of most influence in electron transport calculations. Numerical tests within a density functional theory framework are provided to validate the accuracy and robustness of the proposed method, which in most cases is an order of magnitude faster than conventional methods.

DOI: 10.1103/PhysRevB.77.155301 PACS number(s): 73.40.−c, 73.63.−b, 72.10.−d, 85.65.+h

I. INTRODUCTION

Quantum transport has been an important research subject for more than a decade due to the ever-growing interest in simulating and fabricating nanoscale electronic devices. In particular, the experimental and theoretical investigation of current-voltage (*I-V*) characteristics for molecules and atomic structures placed between conducting electrodes has attracted much effort.[1–11] Most theoretical approaches are based on the Landauer-Büttiker formulation of quantum transport,[12] where the electrical properties of a central interface are described by the transmission coefficients of a number of one-electron states propagating coherently through the system. The widely used Green's function method[13,14] and the wave function matching method[15–17] are two such techniques. To apply these in practice and determine the current through a device under finite bias, it is necessary to evaluate the bulk modes or, correspondingly, the self-energy matrices of each electrode for a considerable number of different energies (chemical potentials) and possibly k points.[18] In many cases, this represents the dominant part of the computational work associated with electron transport calculations, assuming that the Hamiltonian of the system has been provided.

In this paper we develop an efficient method for computing the self-energy matrices using an iterative Krylov subspace technique. The foundation of the method is the evaluation of the self-energy matrices for the semi-infinite electrodes from the solutions of the quadratic eigenvalue problem (QEP) that arises for infinite periodic systems. This approach has been suggested by Ando[19] and studied by several authors.[15,16,20–23] It has been shown[16,24] to be equivalent to well-established iterative and recursive schemes.[25,26] A disadvantage of the latter schemes from a computational point of view is the need to introduce a small imaginary part in the energy in order to ensure that the iterations converge to the correct retarded surface Green's function. This imaginary part forces complex arithmetic in the numerical algorithms used, which is not always the case in the eigenproblem approach.[15,19]

The key motivation for developing the proposed method is the physical observation that only the propagating and the slowly decaying evanescent modes in the bulk electrodes contribute to the transmission of electrons through a semiconductor device of some extension.[8] These modes correspond to the solutions of the QEP that have complex eigenvalues in the vicinity of the unit circle. As recently suggested by Khomyakov *et al.*,[15] this makes it plausible to generate reduced self-energy matrices on the basis of a few selected solutions of the QEP, which include all the electrode-device coupling information that is necessary to determine the correct transmission. To really exploit such an approach in practice, an algorithm to search for and compute *only* the desired quadratic eigenpairs is required.

We will here consider the Arnoldi method[27] combined with a shift-and-invert strategy in order to obtain the QEP solutions. These techniques have proven effective in obtaining selected interior eigenvalues of large-scale general complex eigenproblems.[28–30] In addition, the recent surge of papers studying the Arnoldi procedure applied specifically to polynomial matrix problems indicates that this is a successful technique to build the Krylov subspace for QEPs.[31–34] The algorithm we develop assumes real Hamiltonian matrices (generalization to the complex case is described in Appendix A 2), and targets the complex eigenvalues which are on or inside the unit circle by applying shift-and-invert spectral transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$, where $\hat{\imath}$ is the imaginary unit, and subsequently generating a Krylov subspace for each with the Arnoldi method. Ritz pairs obtained by projecting the QEP onto the three Krylov subspaces give good approximations to the eigenpairs with eigenvalues close to the corresponding shifts. We will show that this method of proceeding is both rigorous and efficient by applying it to various Hamiltonians obtained using density functional theory

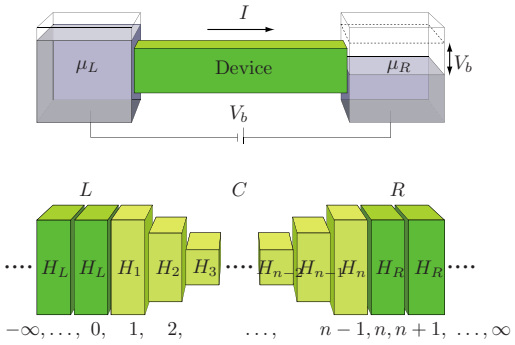

FIG. 1. (Color online) Schematic representation of a two-probe device with applied bias V_b. The top figure illustrates the Landauer-Büttiker picture of coherent scattering between electron reservoirs kept at chemical potentials μ_L and μ_R. The bottom figure shows the device part modeled by two semi-infinite electrodes (L and R) and a central region (C), each divided into principal layers that interact only with nearest-neighbor layers. The layers are described by square Hamiltonian matrices H_i of varying sizes and numbered $i=-\infty,\ldots,\infty$, as indicated.

(DFT) calculations with a localized basis of atomic orbitals.[35]

This paper is organized as follows. In Sec. II we give a brief exposition of our formalism for electron transport. The Krylov subspace method is introduced in Sec. III with details on its key parts: the Arnoldi method, the spectral transformations, and the convergence criterion. Typical convergence behavior is discussed in Sec. IV. The paper ends with numerical examples in Sec. V and a few concluding remarks.

II. ELECTRON TRANSMISSION AND SELF-ENERGY MATRICES

In this section we introduce our formalism, which combines the well-established Green's function method used for electron transport calculations[13,14,36] with the self-energy matrices obtained with the eigenvalue approach of Ando[19] as used in the wave function matching (WFM) method.[15–17] Our goal in combining the methods is to obtain, in the most efficient way, the spectrum of transmission coefficients $T(E)$ for two-probe systems (see top illustration in Fig. 1) in order to calculate the current $I=2e/h\int_{-\infty}^{\infty}T(E)[n_F(E-\mu_L)-n_F(E-\mu_R)]dE$ through the device, where E are the energies, n_F is the Fermi function, and μ_L and μ_R are the chemical potentials of the left (L) and right (R) electron reservoirs.[13,14]

A. Two-probe setup

Consider a two-probe system, as illustrated in the lower part of Fig. 1, where the device corresponds to the central region (C) and the reservoirs are two semi-infinite electrodes (L and R). The system has been divided into principal layers that interact only with nearest-neighbor layers and each layer is assumed to be described by appropriate Hamiltonian $\boldsymbol{H}_i$ and overlap $\boldsymbol{S}_i$ matrices, where i is the layer number, as represented, e.g., in a basis of localized nonorthogonal atomic orbitals. In this manner the Hamiltonian and overlap matrices are block-tridiagonal infinite matrices, where the off-diagonal blocks may be written $\boldsymbol{H}_{i,j}$ and $\boldsymbol{S}_{i,j}$. For the electrode Hamiltonian and overlap matrices we use subscripts L and R instead of numbers i,j. Notice also that the C region in this setup contains at least one layer of each electrode, which means that $\boldsymbol{H}_1=\boldsymbol{H}_L$ and $\boldsymbol{H}_n=\boldsymbol{H}_R$.

We refer the reader to Refs. 13, 14, and 36 for details on how to apply the Green's function method to the current setup. Here we limit ourselves to writing the primary results: First, the finite central region part of the infinite retarded Green's function matrix can be obtained as

$$\boldsymbol{G}_C^r=[(E+\hat{i}\eta)\boldsymbol{S}-\boldsymbol{H}_C-\boldsymbol{\Sigma}_L-\boldsymbol{\Sigma}_R]^{-1}, \tag{1}$$

where η is an infinitesimal quantity, $\boldsymbol{H}_C$ is the central region Hamiltonian, and the effect of the semi-infinite electrodes is accommodated through self-energy matrices $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$. Second, the total transmission coefficient $T(E)$ is then given by

$$T(E)=\mathrm{Tr}\{\boldsymbol{\Gamma}_L\boldsymbol{G}_C^r\boldsymbol{\Gamma}_R\boldsymbol{G}_C^a\}, \tag{2}$$

where $\boldsymbol{\Gamma}_{L/R}=\hat{i}(\boldsymbol{\Sigma}_{L/R}-\boldsymbol{\Sigma}_{L/R}^{\dagger})$ are coupling matrices and $\boldsymbol{G}_C^a$ is the advanced central Green's function matrix, which is obtained from Eq. (1) by using $-\hat{i}\eta$ as the infinitesimal imaginary component in all terms (i.e., implicitly in $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_R$).

We find that an efficient approach (see Appendix A 1) to applying Eqs. (1) and (2) is to compute only a single *diagonal* block of $\boldsymbol{G}_C^r$ in order to evaluate $T(E)$. The question remains how to calculate the required self-energy matrices $\boldsymbol{\Sigma}_{L/R}$ in the most efficient manner.

B. Electrode self-energy matrices from QEPs

It is known that the surface Green's function matrices for a semi-infinite ideal electrode can be evaluated by recursive techniques that take 2^n-1 electrode layers into account in n iterations.[25,26] This is a fast and widely used approach to obtain the self-energy matrices when employing the Green's function method.[1,37]

Another approach has been proposed by Ando,[19] where one constructs and solves an appropriate QEP (introducing notation $\overline{\boldsymbol{H}}\equiv E\boldsymbol{S}-\boldsymbol{H}$)

$$\overline{\boldsymbol{H}}_{L,L}^{\dagger}\phi_k+\lambda_k\overline{\boldsymbol{H}}_L\phi_k+\lambda_k^2\overline{\boldsymbol{H}}_{L,L}\phi_k=0, \tag{3}$$

for $k=1,\ldots,2M_L$, where M_L is the number of orbitals local to the unit cell of the left electrode and similarly for the right electrode with $L\rightarrow R$. The procedure to determine the nontrivial solutions (i.e., the Bloch factors λ_k and electrode modes ϕ_k) from Eq. (3), and subsequently characterize these as propagating or evanescent, right-going (+) or left-going (−), is well described in the literature (we refer the reader to details in Refs. 15 and 16).

Applying Ando's approach via the formalism of the WFM method yields expressions[16]

$$\boldsymbol{\Sigma}_0^L=-\overline{\boldsymbol{H}}_{L,L}^{\dagger}(\boldsymbol{B}_L^-)^{-1}, \tag{4}$$

$$\boldsymbol{\Sigma}_{n+1}^R=-\overline{\boldsymbol{H}}_{R,R}\boldsymbol{B}_R^+ \tag{5}$$

for the electrode self-energy matrices in the layers 0 and n

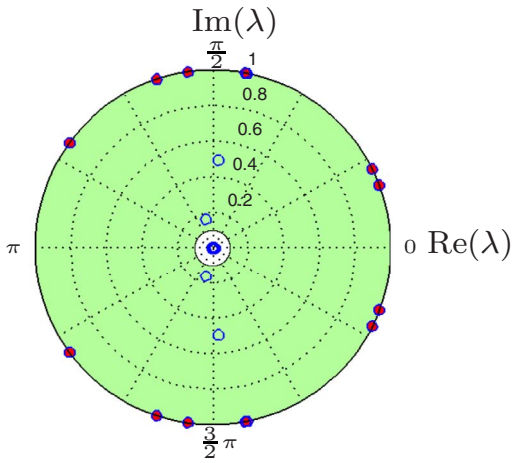

FIG. 2. (Color online) Positions of the 243 complex eigenvalues [blue (circles)] inside the unit disk for a Au(111) electrode with 27 atoms per unit cell at $E=-2$ eV. The 21 eigenvalues corresponding to propagating modes [red (filled) dots] are located on the unit circle. The modes of most significance in transmission calculations are located within the green (shaded) area given by $0.1 \leq |\lambda| \leq 1$.

$+1$ just *outside* the C region, where $\boldsymbol{B}_{L/R}^{\pm}$ are the Bloch matrices constructed from the solutions λ_k and ϕ_k [see the expressions in Ref. 16, in which the notation is $\mathbf{F}_{L/R}(\pm)$ for the Bloch matrices, and $\lambda_n(\pm)$ and $\boldsymbol{u}_n(\pm)$ for the solutions]. After evaluating these self-energy matrices we use them in the Green's function method described above (we set $\eta=0$ in this case, since the retarded Green's function is already uniquely defined by the self-energies[16,21]) and follow the steps outlined in Appendix A 1.

C. Reduced self-energy matrices

From a numerical perspective, it is convenient to keep only those eigenpairs from Eq. (3) that have eigenvalues λ_k within specific intervals[15]

$$\lambda_{\min} \leq |\lambda_k^+| \leq 1, \quad 1 \leq |\lambda_k^-| \leq \lambda_{\min}^{-1}, \tag{6}$$

for a reasonable choice of $\lambda_{\min}$. Evanescent modes with $|\lambda_k|$ outside these intervals are decaying or growing so fast that they have negligible influence in a two-probe setup like ours. The decisive factor in choosing $\lambda_{\min}$ is that the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$ of electrode modes included must be complete in the sense that they can fully represent the transmitted and reflected waves (cf. the WFM formalism).

In what follows, we exploit that a reasonable choice of $\lambda_{\min}$ for transmission calculations with our setup is often of the order 0.1.[38] For example, in the case of the polar plot in Fig. 2, where the Bloch factors with $|\lambda_k| \leq 1$ of a 27-atom Au(111) electrode unit cell are shown, the computationally significant modes can be identified as the eigenvalues inside the shaded area (i.e., by setting $\lambda_{\min}=0.1$). The numerical results given in Sec. V illustrate this observation quantitatively. A proper formal analysis is left for a future publication.[39]

III. KRYLOV SUBSPACE METHOD

In this section, we describe the Krylov subspace method for evaluating the electrode self-energy matrices $\boldsymbol{\Sigma}_0^L$ and $\boldsymbol{\Sigma}_{n+1}^R$. The crucial assumption in the approach is that we may strip the less important modes from the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$, and still obtain a good approximation to the self-energy matrix to be used in transmission calculations. For simplicity, we also assume that the electrode Hamiltonians are real, and give in Appendix A 2 a prescription to generalize to the complex case. Our current method, which targets the specific modes that are most important, can be characterized as a shift-and-invert Arnoldi method with adaptive subspace size. We will describe the key ingredients of the method: the Arnoldi procedure, the spectral transformations, and the convergence criterion. The goal is to present an alternative for obtaining the self-energy matrices, which is faster than existing techniques.

A. Arnoldi procedure

The Krylov subspace of dimension m generated by an $n \times n$ matrix $\boldsymbol{A}$ and an initial vector $\boldsymbol{v}_1$ is given by $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1) \equiv \text{span}\{\boldsymbol{v}_1, \boldsymbol{A}\boldsymbol{v}_1, \boldsymbol{A}^2\boldsymbol{v}_1, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{v}_1\}$.[40] In order to determine this space we apply the Arnoldi procedure[27] which generates an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ for $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$. We use the numerically most stable scheme that employs the modified Gram-Schmidt orthogonalization to successively construct the orthonormal vectors $\boldsymbol{v}_i$. Algorithm I below lists the steps of a continuable version of the Arnoldi procedure which is initially called with a parameter $k=1$ and a random starting vector $\boldsymbol{v}_1$. After $m-1$ iterations the $n \times m$ matrix $\boldsymbol{V}_m=(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ is available.

The projection of the matrix $\boldsymbol{A}$ onto $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$ is then $\boldsymbol{H}_m=\boldsymbol{V}_m^\dagger \boldsymbol{A} \boldsymbol{V}_m$, where $\boldsymbol{H}_m$ is $m \times m$ and upper Hessenberg (i.e., it has zeros below its lower bidiagonal). The matrix $\boldsymbol{H}_m$ is also constructed by Algorithm I. Approximate solutions of the eigenproblem $\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$ can subsequently be obtained as the so-called Ritz eigenpairs $(\gamma, \boldsymbol{V}_m\boldsymbol{y})$ of the projected eigenproblem $\boldsymbol{H}_m\boldsymbol{y}=\gamma\boldsymbol{y}$. As m increases the Ritz pairs become increasingly better approximations to certain eigenpairs of $\boldsymbol{A}$ (we point to Refs. 38 and 39 for details).

Algorithm I: Arnoldi procedure (continuable). Input: $k,m \in \mathbb{Z}$, $\boldsymbol{A} \in \mathbb{R}^{n,n}$, $\boldsymbol{V}_k \in \mathbb{R}^{n,k}$, $\boldsymbol{H}_k \in \mathbb{R}^{k,k}$. Output: $\boldsymbol{V}_{m+1} \in \mathbb{R}^{n,m+1}$, $\boldsymbol{H}_{m+1} \in \mathbb{R}^{m+1,m+1}$.

(1) If $k=1$, $\boldsymbol{v}_1=\boldsymbol{v}_1/\|\boldsymbol{v}_1\|_2$
(2) for $j=k,k+1,\ldots,m$ do
(3) $\boldsymbol{v}=\boldsymbol{A}\boldsymbol{v}_j$
(4) for $i=1,2,\ldots,j$ do
(5) $h_{ij}=\boldsymbol{v}_i^T\boldsymbol{v}$
(6) $\boldsymbol{v}=\boldsymbol{v}-h_{ij}\boldsymbol{v}_i$
(7) end
(8) $h_{j+1,j}=\|\boldsymbol{v}\|_2$
(9) if $h_{j+1,j}=0$, $m=j$, stop
(10) $\boldsymbol{v}_{j+1}=\boldsymbol{v}/h_{j+1,j}$
(11) end

One cannot know in advance how many steps will be needed before the eigenpairs of interest are well approximated by Ritz pairs. If many steps are necessary, then solving the projected eigenvalue problem becomes costly. More-

over, when applying our Krylov method to evaluate the self-energy matrices, we do not know the exact number of eigenpairs wanted and cannot estimate the required dimension of the Krylov subspace.

The first difficulty can be circumvented by restarting the Arnoldi method after a certain number of iterations using the obtained information to generate a better starting vector, or by deflating particular eigenvalues.[41] However, this will not improve on the second difficulty which requires an adaptive maximum dimension of the Krylov subspace. In addition, we observe in most of our applications that the gain from an efficient restart procedure (e.g., as devised by Morgan and Zeng[42]) does not outweigh the computational expense of the restarting overhead. The typical size of the self-energy matrices encountered is too small to make it beneficial to use such techniques, which have been developed for large-scale applications.

Therefore, we have chosen to employ a simple continuation scheme instead of restarting, where a check for convergence is performed after a given number of Arnoldi iterations, and if we are not satisfied, the procedure simply continues where it was left off. With the input parameter k, the listed Arnoldi algorithm is able to generate an initial Krylov subspace $\mathcal{K}_m$ of a given dimension m, but also to continue the process, augmenting the space with subsequent calls. This allows us to perform iterations as long as the approximations are unsatisfactory and/or there is doubt whether all wanted eigenpairs have been found.

An important special case to be considered when applying the Arnoldi procedure to solve an eigenvalue problem is that of algebraically multiple eigenvalues. A Krylov subspace method will, in theory, produce only one eigenvector corresponding to a multiple eigenvalue. So determination of multiplicity is quite difficult. Several approaches exist that deal with this problem, including deflation combined with effects of round-off error,[41] block Arnoldi procedures,[41] and so-called random restarts.[42,43] The present Krylov method does not incorporate any mechanisms to take algebraic multiplicity into account because such cases do not occur in practice for the applications of this work (eigenvalues will not be identical to machine precision in any of the numerical examples, but only to within ~10–11 digits; see Sec. IV).

B. Shift-and-invert transformations

Iterative methods based on Krylov subspaces produce Ritz values that converge fastest to the dominant part of the eigenvalue spectrum given by the extremal eigenvalues.[40] In the current application, it is the interior of the eigenvalue spectrum that is of interest, in particular the eigenvalues λ that satisfy $\lambda_{\min} \leq |\lambda| \leq \lambda_{\min}^{-1}$. To be able to find this part of the spectrum efficiently, we employ a shift-and-invert strategy which implies that the QEP in Eq. (3) is rewritten as

$$(\mu^2 \boldsymbol{M} + \mu \boldsymbol{C} + \boldsymbol{K})\boldsymbol{c}_0 = 0, \tag{7}$$

where

$$\boldsymbol{M} = \overline{\boldsymbol{H}}_{L,L}^T + \sigma \overline{\boldsymbol{H}}_L + \sigma^2 \overline{\boldsymbol{H}}_{L,L}, \tag{8}$$

$$\boldsymbol{C} = \overline{\boldsymbol{H}}_L + 2\sigma \overline{\boldsymbol{H}}_{L,L}, \tag{9}$$

$$\boldsymbol{K} = \overline{\boldsymbol{H}}_{L,L}, \tag{10}$$

and

$$\mu = \frac{1}{\lambda - \sigma}. \tag{11}$$

With this approach, the eigenvalues λ of Eq. (3) have been shifted by σ and inverted while the eigenvectors $\boldsymbol{c}_0$ are unchanged. Thus the dominant part of the spectrum of Eq. (7) now corresponds to the eigenvalues of the original QEP closest to the shift σ.

The simplest and currently state-of-the-art technique for solving Eq. (7) is by linearizing it to a generalized eigenvalue problem of twice the size.[44] In our case $\boldsymbol{M}$ is nonsingular and has size M_L. Therefore, a linearization results in a standard eigenvalue problem of size $2M_L$:

$$\boldsymbol{A}\boldsymbol{x} = \mu \boldsymbol{x}, \tag{12}$$

where $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} 0 & \boldsymbol{I} \\ -\boldsymbol{M}^{-1}\boldsymbol{K} & -\boldsymbol{M}^{-1}\boldsymbol{C} \end{pmatrix}, \tag{13}$$

and the $2M_L$ eigenvalues μ are identical to the ones of Eq. (7). The eigenvectors of Eq. (12) are given by $\boldsymbol{x}^T = (\boldsymbol{c}_0^T, \mu \boldsymbol{c}_0^T)$, so that the original eigenvectors $\boldsymbol{c}_0$ can be selected as the first M_L elements of $\boldsymbol{x}$.

If we assume that the Hamiltonian and overlap matrices for the electrodes are real, then the spectrum of the QEP in Eq. (3) is symmetric with respect to the real axis of the complex plane, and the eigenvalues either are real or occur in complex conjugate pairs.[44] In addition, as seen by transposing Eq. (3), the eigenvalues in this case also come in pairs, λ and $1/\lambda$. We will use these properties to present a simplified method for the extraordinary case of real $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$, and subsequently discuss the steps required for the general complex case in Appendix A 2.

The purpose of the current method is thus to determine the eigenpairs $(\lambda, \boldsymbol{c}_0)$ of Eq. (3) that satisfy $\lambda_{\min} \leq |\lambda| \leq 1$ for a given $\lambda_{\min} > 0$ [the pairs that satisfy $1 \leq |\lambda| \leq \lambda_{\min}^{-1}$ can subsequently be obtained as $(\lambda^{-1}, \boldsymbol{c}_0)$]. As is apparent from the polar plot example in Fig. 2, the majority of the eigenvalues with $|\lambda| \leq 1$ are located near the origin. Therefore, it is not efficient to apply the shift $\sigma = 0$ in order to obtain the wanted eigenvalues, which lie in the outskirts of the unit disk. Instead we may apply four different shifts, given by $\sigma = \pm 1/\sqrt{2}$ and $\sigma = \pm \hat{\imath}/\sqrt{2}$, in four separate Arnoldi procedures. Each of these then covers a quarter slice of the unit disk and produces Ritz values that converge fast to eigenvalues close to the given shift. Simple sorting techniques can be employed in each Arnoldi procedure to take into account only the portion of the Ritz pairs that is covered by a given shift.

When applying the shift-and-invert strategy devised, two of the shifts have to be complex. In practice this means working in complex arithmetic or doubling the size of the problem.[45] However, in the case of real Hamiltonians it is advantageous to search for the complex eigenvalues in con-

jugate pairs and thereby eliminate one of the complex shifts. Moreover, this can be done almost entirely in real arithmetic as follows.

Notice that Eq. (12) was obtained by linearizing the shifted-and-inverted QEP written in Eq. (7). We may also reverse the order of the linearization and shift-and-invert operations. By performing, e.g., a first companion linearization of Eq. (3) that results in an eigenproblem $\hat{\boldsymbol{A}}\boldsymbol{x}=\lambda\boldsymbol{x}$ of double size, and subsequently a shift-and-invert transformation arriving at $(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\boldsymbol{x}=\mu\boldsymbol{x}$, we see that the matrix applied in the Arnoldi procedures can also be written[44]

$$(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}=\begin{pmatrix}-\boldsymbol{M}^{-1}\hat{\boldsymbol{C}} & -\boldsymbol{M}^{-1}\boldsymbol{K}\\ \boldsymbol{I}-\sigma\boldsymbol{M}^{-1}\hat{\boldsymbol{C}} & -\sigma\boldsymbol{M}^{-1}\boldsymbol{K}\end{pmatrix}, \tag{14}$$

where

$$\hat{\boldsymbol{C}}=\overline{\boldsymbol{H}}_L+\sigma\overline{\boldsymbol{H}}_{L,L}. \tag{15}$$

The eigenpairs $(\mu,\boldsymbol{x})$ of $(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\boldsymbol{x}=\mu\boldsymbol{x}$ are exactly the same as those of Eq. (12). In addition, we may now consider the combined spectral transformation for two conjugate shifts σ and σ^*, given by

$$\boldsymbol{T}=(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}(\hat{\boldsymbol{A}}-\sigma^*\boldsymbol{I})^{-1}=\frac{\mathrm{Im}\{(\hat{\boldsymbol{A}}-\sigma\boldsymbol{I})^{-1}\}}{\mathrm{Im}\{\sigma\}}, \tag{16}$$

which was originally proposed by Parlett and Saad.[45] Applying the matrix $\boldsymbol{T}$ in the Arnoldi procedure generates approximate solutions to $\boldsymbol{T}\boldsymbol{x}=\mu'\boldsymbol{x}$, where the eigenvalues are given by

$$\mu'=\frac{1}{(\lambda-\sigma)(\lambda-\sigma^*)}, \tag{17}$$

which becomes extreme for conjugate eigenvalues λ and λ^* of Eq. (3) that are close to σ and σ^*. In our case, the complex shifts are purely imaginary: $\sigma=\hat{\imath}\beta$, where β is real. Then we have $\mu'=(\lambda^2+\beta^2)^{-1}$ and, more importantly, the matrix $\boldsymbol{T}$ is simply given by β^{-1} times the imaginary part of Eq. (14), written as

$$\boldsymbol{T}=\begin{pmatrix}-\beta^{-1}\,\mathrm{Im}\{\boldsymbol{M}^{-1}\hat{\boldsymbol{C}}\} & -\beta^{-1}\mathrm{Im}\{\boldsymbol{M}^{-1}\boldsymbol{K}\}\\ \mathrm{Re}\{\boldsymbol{M}^{-1}\hat{\boldsymbol{C}}\} & \mathrm{Re}\{\boldsymbol{M}^{-1}\boldsymbol{K}\}\end{pmatrix}, \tag{18}$$

which is purely real. This makes it feasible to use real arithmetic in all parts of the algorithm except for the initial complex LU factorization of $\boldsymbol{M}$, which is required for the matrix multiplications by $\boldsymbol{M}^{-1}$.

C. Algorithm and convergence criterion

The algorithm for our Krylov method is composed of two main parts, an iterative part that determines the wanted Ritz pairs $(\lambda,\boldsymbol{c}_0)$ which approximate the eigenpairs of the QEP in Eq. (3), and a noniterative part that sets up the Bloch matrices and evaluates the self-energy matrix from these by direct methods. The iterative part is organized as three independent computations, one for each of the used shifts σ. It consists of the application of the Arnoldi procedure together with a check for convergence plus the initial work to construct the input matrices for Algorithm I. As described in the previous section, the actual calculations will depend on whether the shift is real or imaginary.

The key steps of the Krylov method for evaluating the self-energy matrix $\boldsymbol{\Sigma}^L$ of the left electrode are presented in Algorithm II below. It is important to stress that the details of each step are kept at a minimum to enhance the readability. Furthermore, for evaluating the self-energy matrix $\boldsymbol{\Sigma}^R$ of the right electrode, the steps are exactly the same, except for the substitution $L\rightarrow R$ of all super- and subscripts and the removal of line 1 [this line is only required for left electrodes in order to obtain $\boldsymbol{\Sigma}^L$ from solutions $(\lambda^{-1},\boldsymbol{c}_0)$, e.g., by transposing Eq. (3)]. In the rest of this section we will discuss the main aspects of the algorithm.

Algorithm II: Krylov method to evaluate $\boldsymbol{\Sigma}^L$. Input: $m\in\mathbb{Z}$, $\lambda_{\min}\in[0,1]$, $\overline{\boldsymbol{H}}_L,\overline{\boldsymbol{H}}_{L,L},\overline{\boldsymbol{H}}_{L,L}^T\in\mathbb{R}^{M_L,M_L}$. Output: $\boldsymbol{\Sigma}^L\in\mathbb{C}^{M_L,M_L}$.

(1) Exchange matrices $\overline{\boldsymbol{H}}_{L,L}$ and $\overline{\boldsymbol{H}}_{L,L}^T$

(2) for $\sigma=1/\sqrt{2},-1/\sqrt{2},\hat{\imath}/\sqrt{2}$ do

(3) if σ is real, calculate $\boldsymbol{A}$ from Eq. (13)
else calculate $\boldsymbol{T}$ from Eq. (18) and set $\boldsymbol{A}=\boldsymbol{T}$

(4) select random vector $\boldsymbol{v}_1$ of size $2M_L$

(5) apply Algorithm I to generate $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}_1)$

(6) solve the projected eigenproblem $\boldsymbol{H}_m\boldsymbol{y}=\mu\boldsymbol{y}$

(7) if σ is real, select all $(\mu,\boldsymbol{y})$ that satisfy $\lambda_{\min}\leq|\mu^{-1}+\sigma|\leq 1+\epsilon$, and store the Ritz pairs $(\lambda,\boldsymbol{c}_0)=(\mu^{-1}+\sigma,\boldsymbol{V}_m\boldsymbol{y})$ that have $\mathrm{Re}(\lambda)\mathrm{Re}(\sigma)\geq|\lambda|/2$
else select all $(\mu,\boldsymbol{y})$ that satisfy $\lambda_{\min}\leq|\mu^{-1}+\sigma^2|^{1/2}\leq 1+\epsilon$, and evaluate the eigenvalues λ with the MR-2 method of Ref. 44 and store the Ritz pairs $(\lambda,\boldsymbol{c}_0)=(\lambda,\boldsymbol{V}_m\boldsymbol{y})$ that have $|\mathrm{Im}(\lambda)\mathrm{Im}(\sigma)|>|\lambda|/2$.

(8) for all stored Ritz pairs $(\lambda,\boldsymbol{c}_0)$, find residual $\|(\overline{\boldsymbol{H}}_{L,L}^T+\lambda\overline{\boldsymbol{H}}_L+\lambda^2\overline{\boldsymbol{H}}_{L,L})\boldsymbol{c}_0\|_2$, and check for convergence. If not satisfied, increase m appropriately and go to step 5

(9) end

(10) for all stored Ritz pairs $(\lambda,\boldsymbol{c}_0)$ having $(1+\epsilon)^{-1}\leq\lambda\leq 1+\epsilon$, calculate group velocity v (see Ref. 15); discard the pairs with $v<0$ (i.e., the left-going modes)

(11) evaluate $\boldsymbol{B}_L^+$ and $\boldsymbol{\Sigma}^L=-\overline{\boldsymbol{H}}_{L,L}\boldsymbol{B}_L^+$ from the remaining pairs

First consider the steps 3–8 composing the body of the FOR loop, which are independently executed for the three given shifts σ. Each execution of these steps will determine Ritz pairs that are located in the corresponding quarter-slices of the unit disk. An illustration is shown in Fig. 3 for an Al(100) electrode, where the distinct slices are indicated by shaded areas and the current shifts by crosses. All wanted Ritz pairs found independently for the given shifts are assumed to be collected in a combined set when exiting the loop at step 9.

Initially, in step 3, the linearized and shifted-and-inverted matrix $\boldsymbol{A}$ to be applied in the Arnoldi procedure is determined from Eq. (13) if σ is real and from Eq. (18) if σ is complex. Then a starting vector $\boldsymbol{v}_1$ is selected randomly in step 4. A random starting vector is a reasonable choice in our case, where no prior information about the approximated eigenspace is available. In step 5 the Arnoldi procedure of Algorithm I is called to generate a Krylov subspace of size m, and in step 6, the corresponding eigenpairs $(\mu,\boldsymbol{y})$ of the

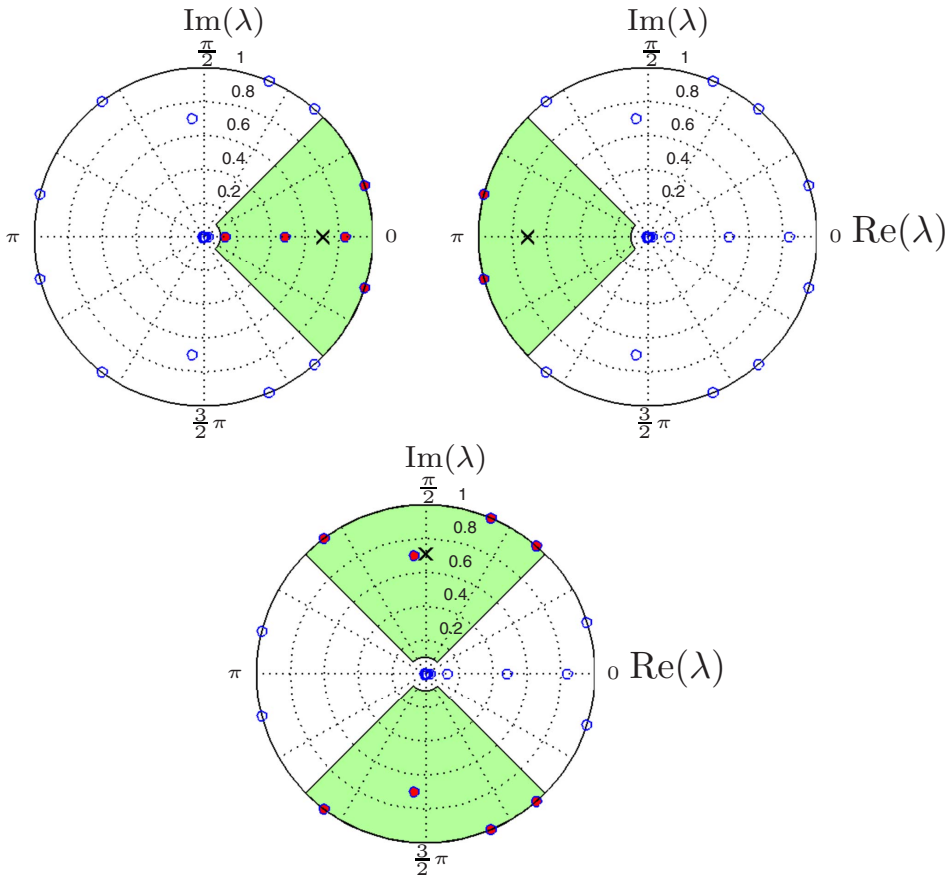

FIG. 3. (Color online) Illustration of the complex eigenvalues [blue (circles)] for the Al(100) electrode at $E=3$ eV. The eigenvalues corresponding to the wanted right-going modes [red (filled) dots] can be separated according to their location within three distinct green (shaded) areas of the unit disk and determined efficiently using shift-and-invert spectral transformations to $\pm 1/\sqrt{2}$ and $\hat{i}/\sqrt{2}$ (crosses).

shifted-and-inverted problem are found by solving the projected eigenproblem with a direct method. This is followed by an elaborate selection scheme to determine which of the available solutions $(\mu,\boldsymbol{y})$ correspond to wanted Ritz pairs $(\lambda,\boldsymbol{c}_0)$ that are located inside the valid quarter slice.

The selection scheme, as outlined in step 7, can be implemented as two separate processes. The first selection process is designed to identify those solutions $(\mu,\boldsymbol{y})$ that correspond to eigenpairs of the original QEP which satisfy $\lambda_{\min} \leq |\lambda| \leq 1$. It is important to realize, however, that, since all computations are done in finite-precision arithmetic, there is no guarantee that the propagating Bloch modes of the electrode will have magnitudes $|\lambda|$ exactly equal to 1. Even the left-going propagating modes that are targeted in our case can have $|\lambda|>1$. In practice, we therefore define the propagating modes to be those Ritz pairs $(\lambda,\boldsymbol{c}_0)$ that satisfy

$$(1+\epsilon)^{-1} \leq |\lambda| \leq 1+\epsilon \tag{19}$$

where ϵ is a small infinitesimal (set to 10^{-8} in our implementation). In order to make sure that all propagating modes are taken into consideration it is thus necessary to select all Ritz pairs that satisfy $\lambda_{\min} \leq |\lambda| \leq 1+\epsilon$.

To obtain the Ritz values λ used in the selection process, we have to transform the solutions $(\mu,\boldsymbol{y})$ of the projected eigenproblem to the corresponding Ritz pairs $(\lambda,\boldsymbol{c}_0)$ by reversing the shift-and-invert operation. The transformation again depends on whether the shift σ is real or imaginary. In the case of real σ, we have $\lambda=\mu^{-1}+\sigma$ from Eq. (11). For imaginary σ, Eq. (17) can be rearranged to $\lambda^2=\mu^{-1}+\sigma^2$, which has two solutions of equal magnitude. This is sufficient to allow selection on the basis of the magnitude $|\lambda|$; however, when it comes to obtaining the Ritz values λ themselves, it is necessary to use other means for imaginary σ, e.g., by forming the Rayleigh quotient.[40] In our case, and for QEPs in particular, it is possible and computationally advantageous to use alternatives to the Rayleigh quotient that work with vectors and matrices of size M_L instead of $2M_L$. Several such techniques that are both fast and accurate have recently been devised by Hochstenbach and van der Vorst.[46] We will adopt the MR-2 method of that paper, which yields $\lambda=\alpha/\beta$, for α and β defined as

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = -\tilde{\boldsymbol{Z}}\overline{\boldsymbol{H}}_{L,L}^{T}\boldsymbol{c}_0, \tag{20}$$

where $\tilde{\boldsymbol{Z}}$ is the pseudoinverse of $\boldsymbol{Z}=(\overline{\boldsymbol{H}}_{L,L}\boldsymbol{c}_0,\overline{\boldsymbol{H}}_{L}\boldsymbol{c}_0)$. Since all eigenvectors are unchanged by the shift-and-invert operation, the $\boldsymbol{c}_0$ vectors applied here are the first M_L elements of the Ritz vectors $\boldsymbol{V}_m\boldsymbol{y}$.

The remaining selection process in step 7 should single out the Ritz pairs that are inside the valid slice of the unit disk. To this end, we can apply the inner product of $(\mathrm{Re}\{\lambda\},\mathrm{Im}\{\lambda\})$ and $(\mathrm{Re}\{\sigma\},\mathrm{Im}\{\sigma\})$, given by

$$\mathrm{Re}\{\lambda\}\mathrm{Re}\{\sigma\}+\mathrm{Im}\{\lambda\}\mathrm{Im}\{\sigma\} = |\lambda||\sigma|\cos\theta, \tag{21}$$

where θ is the angle between λ and σ in a polar representation of the complex plane. In order for λ to be inside the

quarter slice that has σ on the bisector we must have $|\theta| \leq \pi/4$ or equivalently $\cos\theta \geq 1/\sqrt{2}$. For real shifts $\sigma = \pm 1/\sqrt{2}$, this observation yields the condition

$$\frac{\mathrm{Re}\{\lambda\}\mathrm{Re}\{\sigma\}}{|\lambda|} \geq \frac{1}{2}, \tag{22}$$

and similarly for imaginary shift $\sigma = \hat{\imath}/\sqrt{2}$,

$$\frac{|\mathrm{Im}\{\lambda\}\mathrm{Im}\{\sigma\}|}{|\lambda|} > \frac{1}{2}, \tag{23}$$

where the absolute value of the left-hand side is taken to allow λ to be in both the top and the bottom quarter slices. Notice that the equality is removed since the (very rare) event of λ lying exactly on the border of two slices is already taken into account in the condition for real σ.

In step 8 of Algorithm II the check for convergence is carried out. For each shift, the convergence condition is regarded as satisfied when all the Ritz pairs of interest that are also located inside the valid quarter slice are identified and accurate to a given tolerance. We estimate the accuracy of the obtained pairs $(\lambda, \boldsymbol{c}_0)$ by evaluating the corresponding relative residual norm, which yields the following convergence criterion:

$$\frac{\|(\overline{\boldsymbol{H}}_{L,L}^T + \lambda\overline{\boldsymbol{H}}_L + \lambda^2\overline{\boldsymbol{H}}_{L,L})\boldsymbol{c}_0\|_2}{\mathrm{norm}(\overline{\boldsymbol{H}}_L)} \leq \mathrm{tol} \tag{24}$$

where tol is the convergence tolerance and $\mathrm{norm}(\overline{\boldsymbol{H}}_L)$ is an appropriate norm for matrix $\overline{\boldsymbol{H}}_L$. In our implementation we set $\mathrm{tol}=10^{-11}$ and apply the approximation $\mathrm{norm}(\overline{\boldsymbol{H}}_L) \approx \|\mathrm{diag}(\overline{\boldsymbol{H}}_L)\|_2$, that is, we include only the diagonal entries of the two-norm of $\overline{\boldsymbol{H}}_L$. These choices require very low computational effort and give the correct result for all numerical examples we have investigated.

In the event that the convergence check in step 8 of Algorithm II is not satisfied, we assume that the dimension m of the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v}_1)$ generated in step 5, is insufficient. Therefore, we increase m by some fixed amount and go back to step 5 to continue the Arnoldi procedure where it was left off. In the current implementation, we chose to increase the size of the Krylov subspace by $\Delta m = m/2$, where m is the initial value of m given as input. Our experiments show that, for optimal efficiency with this Δm, it is favorable to have the initial m within the range 30–50 if the sizes of the input matrices are of order less than 1000. After convergence has been achieved, the final steps 10–11 of Algorithm II present the operations required to collect the Ritz pairs that have been determined and subsequently obtain the self-energy matrix.

IV. TYPICAL CONVERGENCE BEHAVIOR

In this section, we briefly exemplify the typical convergence behavior of Algorithm II by monitoring the relative residual norm of the wanted eigenpairs as a function of the number of iterations. An expression for this norm for a given eigenpair $(\lambda, \boldsymbol{c}_0)$ is available as the left-hand side of Eq. (24). We will consider the Al(100) electrode at $E=3$ eV and parameter $\lambda_{\min}=0.1$, which requires a total of 13 eigenpairs to be determined (eight propagating modes and five evanescent modes) from the three separate Arnoldi procedures. This example corresponds to the situation illustrated in Fig. 3 and represents a typical calculation for an Al(100) electrode with 18 atoms per unit cell (the size of the self-energy matrix is 72).

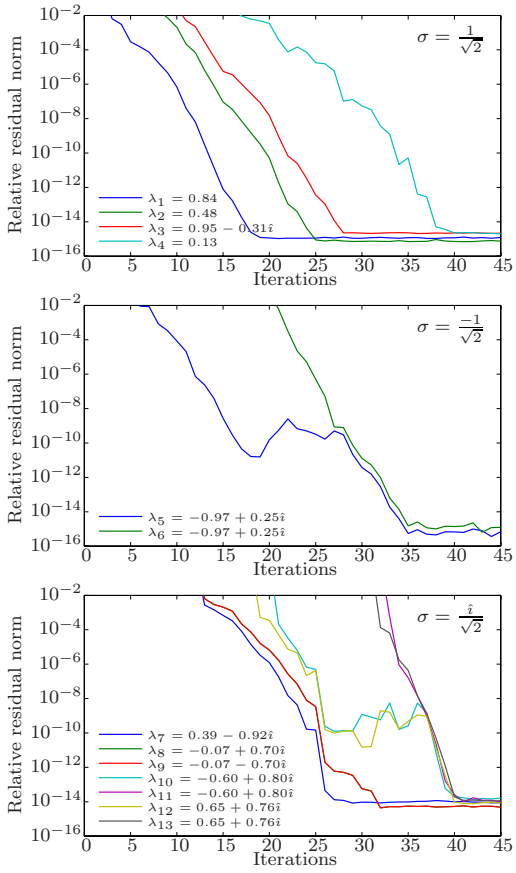

FIG. 4. (Color online) Convergence behavior of the Krylov algorithm for the Al(100) electrode at $E=3$ eV. The figures show the residual norm as a function of iterations for Ritz pairs that satisfy $0.1 \leq |\lambda| \leq 1+\epsilon$, in the case of shift-and-invert transformations to $\pm 1/\sqrt{2}$ and $\hat{\imath}/\sqrt{2}$, respectively.

In Fig. 4 we present curves showing the history of the residual norms for the wanted eigenpairs in each of the separate shift-and-invert Arnoldi procedures. We show only the 45 first iterations since this number is enough for convergence in all cases. Also, only residuals for eigenpairs corresponding to right-going modes are displayed.

The top figure of Fig. 4 illustrates the results from applying the shift $\sigma = 1/\sqrt{2}$ and shows that the Arnoldi procedure determines four different Ritz pairs with individual convergence curves. Comparing with the corresponding polar plot in Fig. 3 (top left), we observe a fifth eigenvalue $(\lambda = 0.95 + 0.31\hat{\imath})$ located inside the valid quarter slice. This fifth eigenvalue represents a left-going mode and is thus discarded in step 10 of Algorithm II. We also see by comparison with Fig. 3 that the eigenpair with eigenvalues furthest from the current shift (the cross) in the complex plane, in this case λ_4, is the slowest to converge.

The middle figure of Fig. 4 shows the convergence of the two Ritz pairs that are covered by the Arnoldi procedure with $\sigma=-1/\sqrt{2}$ and correspond to right-going modes in the present example. We note that λ_5 and λ_6 are nearly multiple eigenvalues, and that the behavior of the residual norms, where one eigenpair is available many iterations before its counterpart, is typical in such a case. Here, in particular, we see that eigenvalue λ_5 is determined to an accuracy of $\sim 10^{-11}$ after 18 iterations before λ_6 even shows up as a Ritz value of the projected eigenproblem. This indicates that λ_5 and λ_6 must be identical to around ten significant digits, and that they cannot be distinguished in our Arnoldi procedure before this accuracy is achieved. Without additional mechanisms to deal with multiple eigenvalues this then implies an upper bound condition on the value of the tol parameter.

The bottom figure of Fig. 4 shows the residual norm history of the remaining seven Ritz pairs required in the current example. These are determined by the Arnoldi procedure with imaginary shift $\sigma=\hat{i}/\sqrt{2}$ and correspond to filled dots in the bottom polar plot of Fig. 3 which represent right-going modes. We observe that the eigenvalue closest to σ, here denoted by λ_8, constitutes a complex conjugate pair together with λ_9, and that these have exactly the same residual norm curve (indistinguishable in the figure), although they are obtained separately as individual Ritz pairs in the algorithm.

In all residual norm figures, we see the trend that the eigenvalues located far from the position of the shift are slow to converge. This suggests that eigenvalues located in the vicinity of the intersections between the unit circle and the dividing lines of the four quarter slices will be the most difficult to determine since they are furthest from the corresponding shifts. The maximum distance from such an eigenvalue to σ is $1/\sqrt{2}$, which is the same as from σ to the origin. This raises concern whether the many unwanted eigenvalues close to the origin can become dominant compared to the wanted border eigenvalues. Fortunately, this is not the case because the unwanted eigenvalues close to the origin are clustered and therefore easy to represent in the Krylov subspace with only a few iterations.[40] We observe this in practice, e.g., from the bottom figure of Fig. 4, where the Ritz pair corresponding to λ_{12}, which lies close to the worst-case position on the unit circle, initially converges only slightly slower than the Ritz pair for λ_8 positioned right next to the shift.

V. NUMERICAL EXAMPLES

To illustrate the accuracy and practical aspects of the proposed Krylov subspace method we present transmission calculations for a metal-device-metal system that has been widely studied in the literature. In addition, we compute the current through this system at 1 and 2 V biases, and use the parameter $\lambda_{\min}$ to investigate the significance of the evanescent modes in obtaining the correct currents. Last, we apply the method to evaluate the self-energy matrices of a variety of electrodes (different types and sizes) and compare the actual measured CPU times[47] with those required by conventional methods.

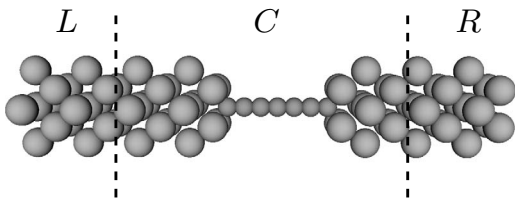

FIG. 5. Schematic illustration of the Al(100)-C7-Al(100) two-probe system.

A. Carbon wire between aluminum electrodes

To demonstrate the applicability of the proposed Krylov subspace method, we first consider carbon chains coupled to metallic electrodes, which have been investigated in detail recently.[1,5,6] Carbon atomic wires are interesting conductors since the equilibrium conductance of short monatomic chains varies with their length in an oscillatory fashion. We will examine the two-probe system shown in Fig. 5 corresponding to a straight wire of seven carbon atoms attached to Al(100) electrodes (lattice constant 4.05 Å). This structure exhibits a local maximum in the oscillatory conductance since it represents an odd-numbered C chain.[5] In our configuration, we fix the C-C distance to 2.5 bohrs and the distance between the ends of the carbon chain and the first plane of Al atoms at 1.0 Å. We use single-ζ basis sets for both types of atoms. The considered Al(100) electrode unit cell consists of 18 atoms in four layers with identical unit cells for the left and right electrodes. Notice that we do not use any symmetry properties of the metallic electrode to reduce the lateral size of the cells (as done, e.g., in Ref. 17) but rather use the full size matrices in Algorithm II. The same system has been studied by Brandbyge *et al.*[1]

We apply the proposed Krylov subspace method to calculate the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ of the left and right electrodes for a range of energies $E\in[-4\ \text{eV}, 4\ \text{eV}]$ and for different choices of the parameter $\lambda_{\min}$. The self-energy matrices are then used in the evaluation of the corresponding transmission coefficients $T(E)$.

Figure 6 presents the results for bias voltages $V_b=0$, 1, and 2 V in three cases of $\lambda_{\min}$. These significant bias settings are chosen for benchmarking and comparison reasons. The (black) full curves corresponding to $\lambda_{\min}=0.1$ reproduce the transmission spectra obtained in Ref. 1 (for 0 and 1 V) exactly except for the peak at $E=3.63$ eV (for 0 V), which is probably due to finer sampling in our work. In addition, we have calculated the similar curve with the full sets of electrode modes and the results are indistinguishable from those with the setting $\lambda_{\min}=0.1$ (and therefore not displayed in the figure). We note this as quantitative verification that the exclusion of the rapidly decaying evanescent modes is plausible in our setup.

We also see in Fig. 6 that the curves for the parameter $\lambda_{\min}$ set to 0.1 [black (full)] and 0.5 [red (dashed)] are almost identical, which indicates that the vast majority of the evanescent modes (those satisfying $|\lambda|<0.5$) have very little influence on $T(E)$ in the energy regime considered. However, when $\lambda_{\min}$ is set to 0.99 [blue (dotted curves)], in which case only propagating modes and very close to propagating modes are included in the evaluation of self-energy matrices,

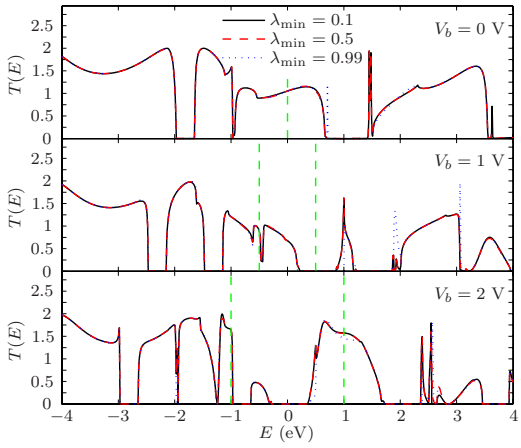

FIG. 6. (Color online) Transmission spectrum of the Al(100)-C7-Al(100) system for different bias voltages V_b. The self-energy matrices used in the $T(E)$ calculations have been obtained at the Γ point by the proposed Krylov subspace method with parameter $\lambda_{\min}$ at several settings: 0.1 [black (full) curve], 0.5 [red (dashed) curve], and 0.99 [blue (dotted) curve]. The bias windows are indicated by the vertical dashed lines.

there are several noticeable deviations from the other curves. Also inside the bias windows and especially for V_b=2 V, the disregard of the evanescent modes produces errors in the obtained transmission coefficients $T(E)$.

The deviations become even more evident in Fig. 7, where the current is displayed as a function of the parameter $\lambda_{\min}$ for nonzero bias voltages. As the value of $\lambda_{\min}$ is increased from around 0.5 to 1, the computed current I starts to

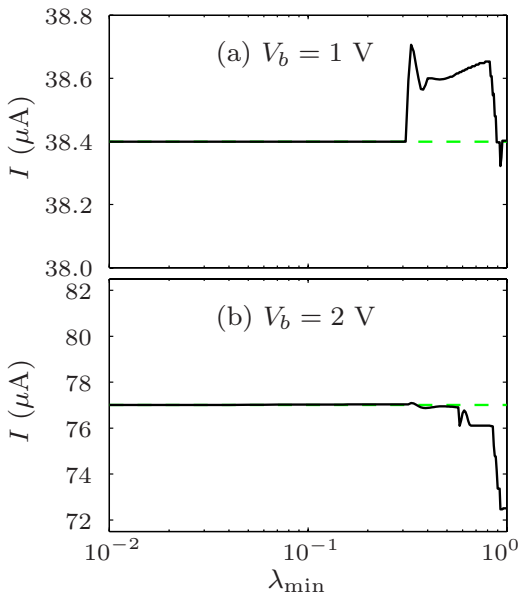

FIG. 7. (Color online) Current as a function of the parameter $\lambda_{\min}$ used by the Krylov subspace method for the Al(100)-C7-Al(100) system with applied bias voltages V_b= (a) 1 and (b) 2 V. The correct currents obtained by conventional methods are $I \approx 38.4$ and ≈ 77.0 μA, respectively, indicated here by the green (dashed) lines.

TABLE I. CPU times in seconds for computing the left self-energy matrix $\mathbf{\Sigma}_L$ at 20 different energies E between −2 and 2 eV for selected electrode types and matrix sizes N. The parameter $\lambda_{\min}$ was set to 0.1.

Electrode type	Size	2^n iterative	DGEEV	Krylov
Li[a]	16	0.1	0.0	0.0
Fe[b]	54	4.2	2.3	0.6
Al(100)[c]	72	4.9	3.3	0.8
Al(100)[c]	128	27.9	17.5	3.6
Au(111)[d]	243	167.2	73.7	11.5
(2,2) CNT[e]	64	3.6	2.4	0.7
(4,4) CNT[e]	128	26.0	14.4	2.9
(8,8) CNT[e]	256	208.8	118.8	17.0
(12,12) CNT[e]	384	608.4	373.6	45.6
(16,16) CNT[e]	512	1230.0	1403.9	121.5
(20,20) CNT[e]	640	1542.3	1125.7	148.0

[a]Measurements from transmission calculations for ideal Li system.
[b]Measurements from transmission calculations for Fe-MgO-Fe; see geometry description in Ref. 10.
[c]Measurements from transmission calculations for Al(100)-C7-Al(100) described in this work (see also Ref. 1).
[d]Measurements from transmission calculations for Au(111)-BDT-Au(111); see, e.g., description in Ref. 11.
[e]Measurements from transmission calculations for ideal armchair (n,n) carbon nanotubes; see, e.g., description in Ref. 4.

depart significantly from the correct value. Therefore, we anticipate that at least some slowly decaying evanescent modes must be taken into account in order to describe the transmission properties of the Al(100)-C7-Al(100) system. Moreover, we see that this can be achieved in a rigorous and systematic fashion by selecting $\lambda_{\min}$ appropriately when using the proposed Krylov subspace method to calculate the self-energy matrices.

B. CPU run times

In this section we focus on the typical savings in the computational time that can be achieved when computing the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ with the proposed Krylov subspace method. We will compare run times directly with some conventional schemes usually applied in electron transport calculations. Our aim is to illustrate a significant speedup in calculating the self-energy matrices. This is of interest in future efforts to model much larger systems, and, in particular, for electrode unit cells that do not have any lateral symmetry properties.

Table I presents the profiling results when applying three different methods to calculate the same left self-energy matrix $\mathbf{\Sigma}_L$ for common types of electrodes and various matrix sizes N. In every case we consider only the Γ point and use single-ζ basis sets, except for Au(111) where a double-ζ-polarized set is used. Since the computational cost can vary significantly with E, the seconds listed represent the accumulated time of 20 independent calculations at equidistant energies in the interval $E \in [-2\ \text{eV}, 2\ \text{eV}]$. We focus on the

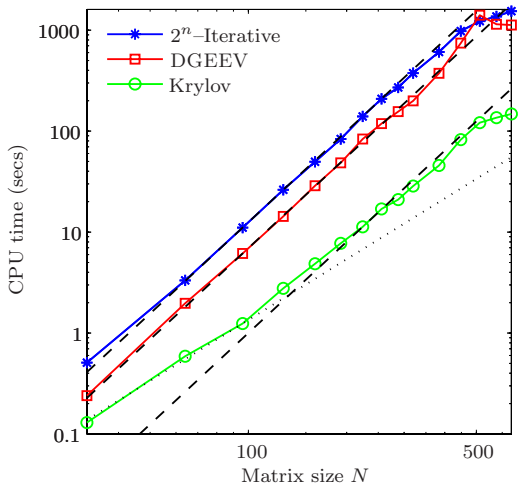

FIG. 8. (Color online) CPU times for computing the left self-energy matrix $\mathbf{\Sigma}_L$ plotted as a function of the size N of $\mathbf{\Sigma}_L$ for a range of armchair (n,n) CNT electrodes, where $n=1,\ldots,20$. The dotted and dashed lines indicate $O(N^2)$ and $O(N^3)$ computational complexity, respectively.

profiling for general electrode configurations and do not use lattice symmetries to reduce the order of the unit cells to elementary size even when this is possible.[17]

In the third column of Table I the run times to compute the correct self-energy matrices with the widely used iterative scheme of López Sancho *et al.*[26] are displayed. As the error in $\mathbf{\Sigma}_L$ obtained by this technique is reduced by $1/2^n$ after n iterations (we denote this method as 2^n iterative), it generally converges in $n\sim 22$ steps. In addition, run times for the conventional eigenvalue approach to evaluating the self-energy matrices, in which a standard eigensolver is used to determine the full set of modes, are presented in the fourth column. For this version, we simply substituted part of our Krylov subspace algorithm (steps 1–9 of Algorithm II) with the state-of-the-art LAPACK routine DGEEV.[47] In the last column the time required by the proposed Krylov subspace method is shown. In all cases of the latter the parameter $\lambda_{\min}$ was set to 0.1.

From the profiling results in Table I we see that the computational time of the Krylov subspace method is significantly reduced compared with the presently widely used 2^n-iterative technique. Also the conventional eigensolver scheme using DGEEV is typically faster than the 2^n-iterative algorithm [the exception for the (16,16) carbon nanotube (CNT) is related to cache usage[48]]. A comparison of the timings in the last two columns verifies that the cost to evaluate the self-energy matrices from only the few most important modes of the electrodes, as in our Krylov subspace method, is in general much lower than required by a direct eigensolver to determine all possible modes.

In order to illustrate the computational complexity of the methods we show the CNT run times as a function of the matrix size N in a logarithmic plot in Fig. 8. Clearly, all methods have $O(N^3)$ complexity; however, the Krylov subspace method initially follows the typical $O(N^2)$ complexity of the Arnoldi procedure[49] until the cost of the shift-and-invert operations becomes dominant. For $N>500$ we observe effects due to more and sometimes less favorable cache usage. Overall, we see that the Krylov subspace method is fastest by an order of magnitude for all but the smallest cases.

It is important to point out that the obtained self-energy matrices $\mathbf{\Sigma}_L$ are in all cases applied in a subsequent transmission calculation of $T(E)$ for the two-probe systems indicated in Table I, and the results then checked against those of the conventional methods [the resulting transmissions $T(E)$ are identical for the three methods in all cases of E to at least three decimals]. Furthermore, the setting of the parameter $\lambda_{\min}$ to 0.1 yields self-energy matrices evaluated from all the modes that have phases λ satisfying $0.1<|\lambda|<1+\epsilon$. This is more than adequate for obtaining correct results to an accuracy of three decimals for all the systems considered in this section. In practice, the parameter $\lambda_{\min}$ can often be selected >0.1 if lower accuracy in the $T(E)$ calculation is satisfactory, and this would show off the approach as even faster.

VI. CONCLUSIONS

In conclusion, we have developed an efficient and robust Krylov subspace method for evaluating the self-energy matrices that are required in electron transport calculations of nanoscale devices. The method exploits the observation that only the propagating and slowly decaying evanescent modes in the electrodes are computationally significant for determining the transmission coefficients when the system is appropriately set up.

The proposed method is based on the Arnoldi procedure and applies carefully chosen shift-and-invert spectral transformations to enhance the convergence toward the wanted interior eigenpairs that correspond to significant modes. We have investigated the convergence properties and shown that the accuracy and efficiency are mainly controlled by two parameters: the tolerance tol to be satisfied by of the relative residuals of the obtained Ritz values and the parameter $\lambda_{\min}$ that implicitly sets the number of modes taken into account.

In Sec. V we tested the Krylov subspace method on a metal-device-metal system and compared it to conventional methods. The applications show that the proposed method can be applied to calculate the transmission characteristics in a rigorous and systematic fashion and that the basic assumption of only including selective solutions in the electrode self-energy matrix is valid for many two-probe systems. The overall saving in computational time achieved by the Krylov subspace method is significant and in most cases more than an order of magnitude in comparison with conventional methods.

ACKNOWLEDGMENTS

The authors would like to thank J. Taylor and the people at Atomistix for helpful discussions. This work was supported by the Danish Council for Strategic Research (NABIIT) under Grant No. 2106-04-0017, "Parallel Algorithms for Computational Nano-Science."

APPENDIX A: COMPUTATIONAL DETAILS

1. Fast transmission calculation

We give the numerical steps to efficiently evaluate $T(E)$ via Eqs. (1) and (2). From the outset, the computational costs are reduced by taking into account that the self-energy matrices are nonzero only in the corner blocks, that is,

$$\boldsymbol{G}_C = \begin{pmatrix} \overline{\boldsymbol{H}}_1 - \boldsymbol{\Sigma}_1^L & \overline{\boldsymbol{H}}_{1,2} & & & \\ \overline{\boldsymbol{H}}_{1,2}^\dagger & \overline{\boldsymbol{H}}_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \overline{\boldsymbol{H}}_{n-1} & \overline{\boldsymbol{H}}_{n-1,n} \\ & & & \overline{\boldsymbol{H}}_{n-1,n}^\dagger & \overline{\boldsymbol{H}}_n - \boldsymbol{\Sigma}_n^R \end{pmatrix}^{-1}, \tag{A1}$$

where the self-energy blocks are numbered similarly to the Hamiltonian blocks. We then select a given diagonal block k and define self-energy matrices for every layer of the system, as[50–52]

$$\boldsymbol{\Sigma}_i^L = \overline{\boldsymbol{H}}_{i-1,i}^\dagger (\overline{\boldsymbol{H}}_{i-1} - \boldsymbol{\Sigma}_{i-1}^L)^{-1} \overline{\boldsymbol{H}}_{i-1,i}, \quad -\infty < i \le k, \tag{A2}$$

$$\boldsymbol{\Sigma}_i^R = \overline{\boldsymbol{H}}_{i,i+1} (\overline{\boldsymbol{H}}_{i+1} - \boldsymbol{\Sigma}_{i+1}^R)^{-1} \overline{\boldsymbol{H}}_{i,i+1}^\dagger, \quad k \le i < \infty, \tag{A3}$$

which can be used to recursively evaluate the self-energy matrices $\boldsymbol{\Sigma}_k^L$ and $\boldsymbol{\Sigma}_k^R$ when the matrices $\boldsymbol{\Sigma}_1^L$ and $\boldsymbol{\Sigma}_n^R$ (or $\boldsymbol{\Sigma}_0^L$ and $\boldsymbol{\Sigma}_{n+1}^R$ of the semi-infinite electrodes) are available. The kth block of the Green's function matrix is now given by

$$\boldsymbol{G}_{k,k} = (\overline{\boldsymbol{H}}_k - \boldsymbol{\Sigma}_k^L - \boldsymbol{\Sigma}_k^R)^{-1}, \tag{A4}$$

which corresponds to inverting the block of smallest size in the system, if k is chosen accordingly. Finally Eq. (2) is applied in a simplified version

$$T(E) = \mathrm{Tr}\{\boldsymbol{\Gamma}_k^L \boldsymbol{G}_{k,k} \boldsymbol{\Gamma}_k^R \boldsymbol{G}_{k,k}^\dagger\}, \tag{A5}$$

where the relation $\boldsymbol{G}_{k,k}^a = (\boldsymbol{G}_{k,k}^r)^\dagger$ between the advanced (a) and retarded (r) Green's functions is used [$\boldsymbol{G}^a = (\boldsymbol{G}^r)^\dagger$ is valid when E is real, since $\boldsymbol{H}$ is Hermitian and $\boldsymbol{\Sigma}^a = (\boldsymbol{\Sigma}^r)^\dagger$; see Ref. 13].

2. Generalization to complex Hamiltonian matrices and k-point sampling

In the Krylov subspace method presented in this paper we have assumed that the electrode Hamiltonian matrices are real in order to simplify the computational procedures. We now discuss the steps required to handle the case of complex $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$, which is the case, e.g., when applying k-point sampling (Algorithm II works only for the Γ point).

As noted in Sec. III B, the assumption of real $\overline{\boldsymbol{H}}_L$ and $\overline{\boldsymbol{H}}_{L,L}$ leads to simplifications with the shift-and-invert operations:

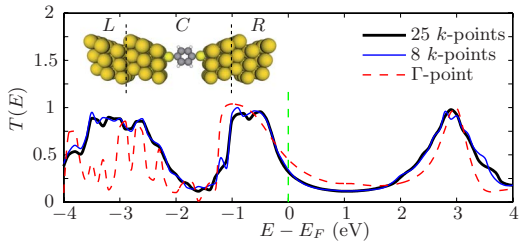

FIG. 9. (Color online) Transmission spectrum of the Au(111)-BDT-Au(111) system for different k-point samplings and V_b=0. The self-energy matrices used in the $T(E)$ calculations have been obtained by the generalized Krylov subspace method with parameter $\delta_{\min}$=0.1.

First, we may consider only right-going modes $(\lambda, \boldsymbol{c}_0)$ with $|\lambda| \le 1$ since the left-going modes are uniquely related as $(\lambda^{-1}, \boldsymbol{c}_0)$, and, second, we can use the spectral transformation $\boldsymbol{T}$ in Eq. (18) to determine the wanted eigenpairs for the two imaginary shifts $\sigma = \pm \hat{\imath}/\sqrt{2}$ simultaneously and in real arithmetic.

In order to generalize the Krylov subspace method to complex Hamiltonian matrices, it is thus necessary to determine the left-going modes satisfying $1 \le |\lambda| \le \lambda_{\min}^{-1}$ (i.e, located outside the unit circle) directly, since there is no general relation to the right-going modes (we note that it is advantageous to change the shift positions to be outside the unit circle, although this is not necessary for good convergence). Furthermore, we must abandon the $\boldsymbol{T}$ matrix and perform two independent shift-and-invert operations for $\sigma = \pm \hat{\imath}/\sqrt{2}$. It is clear that all this is now done in complex arithmetic and that the extra shift required will make the general algorithm a little more expensive (as shown in Sec. V B, the LU factorization required for each shift-and-invert operation is the dominant cost of our approach).

We have implemented the generalization and can illustrate its applicability by converging the transmission spectrum of the benzene di-thiol (BDT) molecule coupled to gold (111) surfaces in Fig. 9 by 3×3 and 7×7 k-point sampling of the Monkhorst type.[53] The calculation setup used is exactly the same as in Ref. 11 and the results can be confirmed.[3,11] Also, we have computed $T(E)$ for each E and k with self-energy matrices of both the 2^n-iterative method and the Krylov subspace method and checked that the results are identical to within three decimals. The CPU times required for, e.g., the 3×3 curve (eight k points) were 167 and 32 min for the two methods, respectively, while the Γ-point curve takes 2.7 min with Algorithm II. We conclude that the generalized Krylov subspace algorithm is, in this case, 1.5 times slower (per k point) than the real matrix version presented in Sec. III but still more than five times faster than the commonly used 2^n-iterative approach.

*hhs@imm.dtu.dk

[1] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, Phys. Rev. B **65**, 165401 (2002).

[2] M. Di Ventra, S. T. Pantelides, and N. D. Lang, Phys. Rev. Lett. **84**, 979 (2000).

[3] S. V. Faleev, F. Léonard, D. A. Stewart, and M. van Schilfgaarde, Phys. Rev. B **71**, 195422 (2005).

[4] H. S. Gokturk, in *Proceedings of the Fifth IEEE Conference on Nanotechnology*, 2005, Vol. 2, pp. 677–680.

[5] N. D. Lang and P. Avouris, Phys. Rev. Lett. **84**, 358 (2000).

[6] B. Larade, J. Taylor, H. Mehrez, and H. Guo, Phys. Rev. B **64**, 075420 (2001).

[7] A. Nitzan and M. A. Ratner, Science **300**, 1384 (2003).

[8] P. Pomorski, C. Roland, and H. Guo, Phys. Rev. B **70**, 115408 (2004).

[9] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, Science **278**, 252 (1997).

[10] M. Stilling, K. Stokbro, and K. Flensberg, Mol. Simul. **33**, 557 (2007).

[11] K. Stokbro, J.-L. Mozos, P. Ordejon, M. Brandbyge, and J. Taylor, Comput. Mater. Sci. **27**, 151 (2003).

[12] M. Büttiker, Y. Imry, R. Landauer, and S. Pinhas, Phys. Rev. B **31**, 6207 (1985).

[13] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge U.K., 1995).

[14] Y. Meir and N. S. Wingreen, Phys. Rev. Lett. **68**, 2512 (1992).

[15] P. A. Khomyakov and G. Brocks, Phys. Rev. B **70**, 195402 (2004).

[16] P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. J. Kelly, Phys. Rev. B **72**, 035450 (2005).

[17] K. Xia, M. Zwierzycki, M. Talanana, P. J. Kelly, and G. E. W. Bauer, Phys. Rev. B **73**, 064420 (2006).

[18] K. S. Thygesen and K. W. Jacobsen, Phys. Rev. B **72**, 033401 (2005).

[19] T. Ando, Phys. Rev. B **44**, 8017 (1991).

[20] P. S. Krstić, X.-G. Zhang, and W. H. Butler, Phys. Rev. B **66**, 205319 (2002).

[21] D. H. Lee and J. D. Joannopoulos, Phys. Rev. B **23**, 4997 (1981).

[22] S. Sanvito, C. J. Lambert, J. H. Jefferson, and A. M. Bratkovsky, Phys. Rev. B **59**, 11936 (1999).

[23] T. Shimazaki, H. Maruyama, Y. Asai, and K. Yamashita, J. Chem. Phys. **123**, 164111 (2005).

[24] J. Velev and W. Butler, J. Phys.: Condens. Matter **16**, R637 (2004).

[25] F. Guinea, C. Tejedor, F. Flores, and E. Louis, Phys. Rev. B **28**, 4397 (1983).

[26] M. P. Lopez Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio, J. Phys. F: Met. Phys. **15**, 851 (1985).

[27] W. E. Arnoldi, Q. Appl. Math. **9**, 17 (1951).

[28] M. N. Kooper, H. A. van der Vorst, S. Poedts, and J. P. Goedbloed, J. Comput. Phys. **118**, 320 (1995).

[29] K. Meerbergen and D. Roose, IMA J. Numer. Anal. **16**, 297 (1996).

[30] N. Nayar and J. M. Ortega, J. Comput. Phys. **108**, 8 (1993).

[31] Z. Bai and Y. Su, SIAM J. Matrix Anal. Appl. **26**, 640 (2005).

[32] L. Hoffnung, R.-C. Li, and Q. Ye, Linear Algebr. Appl. **415**, 52 (2006).

[33] U. B. Holz, G. H. Golub, and K. H. Law, SIAM J. Matrix Anal. Appl. **26**, 498 (2004).

[34] Q. Ye, Appl. Math. Comput. **172**, 818 (2006).

[35] First-principles DFT calculations are done with the commercial software package ATOMISTIX TOOLKIT 2.0. We use norm-conserved pseudopotentials for the core electrons and the local density approximation for the exchange-correlation potential (Ref. 1). More details about the software can be found on the company website (www.atomistix.com).

[36] P. N. C. Caroli, R. Combescot, and D. Saint-James, J. Phys. C **4**, 916 (1971).

[37] M. B. Nardelli, Phys. Rev. B **60**, 7828 (1999).

[38] A brief explanation for this is that, since the boundary layers of the C region in our setup are given by principal electrode layers, the evanescent modes that decay very fast do not "survive" the propagation through these layers and therefore do not give any components outside the sets $\{\phi_k^+\}$ and $\{\phi_k^-\}$ at the boundaries of C.

[39] H. H. B. Sørensen, D. E. Petersen, S. Skelboe, P. C. Hansen, and K. Stokbro (unpublished).

[40] L. N. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997).

[41] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide* (SIAM, Philadelphia, 2000).

[42] Ronald B. Morgan and M. Zeng, Linear Algebr. Appl. **415**, 96 (2006).

[43] Z. Jia, J. Comput. Math. **17**, 257 (1999).

[44] F. Tisseur and K. Meerbergen, SIAM Rev. **43**, 235 (2001).

[45] B. N. Parlett and Y. Saad, Linear Algebr. Appl. **88–89**, 575 (1987).

[46] M. E. Hochstenbach and H. A. van der Vorst, SIAM J. Sci. Comput. **25**, 591 (2003).

[47] All computations in this work were done on a Sun ULTRASPARC IV dual-core CPUs (1350 MHz/8 MB L2-cache). We use the vendor-supplied Sun Performance Library that includes platform-optimized versions of LAPACK routines.

[48] For the armchair (16,16) CNT electrode (N=512) the call to DGEEV produces an extremely high number of L2 cache misses, many more than for the larger (18,18) CNT electrode (N=576). This causes the very poor run times of the DGEEVmethod for this particular electrode.

[49] G. W. Stewart, *Matrix Algorithms* (SIAM, Philadelphia, 2001).

[50] E. M. Godfrin, J. Phys.: Condens. Matter **3**, 7843 (1991).

[51] D. E. Petersen, H. H. B. Sørensen, S. Skelboe, P. C. Hansen, and K. Stokbro, J. Comput. Phys. **227**, 3174 (2008).

[52] S. Y. Wu, J. Cocks, and C. S. Jayanthi, Phys. Rev. B **49**, 7957 (1994).

[53] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).